Genetic Algorithm Based Clustering Optimization

A Survey

Rawaa Nadhum Saeed
Dept. of computer science
University of Babylon, College Science for Women
Babylon, Iraq
Rawaa.saeed.gsci8@student.uobabylon.edu.iq
https://orcid.org/0009-0004-8022-289X?lang=en

Mahdi Abed Salman
Dept. of computer science
University of Babylon, College Science for Women
Babylon, Iraq
Mahdi.salman@uobabylon.edu.iq
https://orcid.org/0000-0002-7805-6800

Muhammed Abaid Mahdi
Dept. of computer science
University of Babylon, College Science for Women
Babylon, Iraq
Wsci.muhammed.a@uobabylon.edu.iq
https://orcid.org/0000-0002-7820-4340

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100105

Received Jun. 6, 2022. Accepted for publication Aug. 10, 2022

***Abstract*—*Genetic algorithm has an important role in improving clustering methods. When it comes to getting into a dataset, one of the most common methods used is clustering. Recent years have seen a rise in the number of articles interested in clustering, which may be attributed to the development of various new fields of application. Most clustering method's performance depends on initial values of parameters or the preprocessing of a dataset, so clustering needs to optimize. The improvement of clustering in two directions, either we improve the parameters or improve the input data. Ga is an evolutionary technique well-known in optimization. This survey deals with the methods that use a genetic algorithm to choose parameter values and input data. This study shows many important performance metrics used to get the optimal result in each research.*

Keywords—Genetic Algorithm, Clustering Algorithms, Optimization, GA based Clustering

I. INTRODUCTION

Clustering is a data analysis process applied to classify the unlabeled data [1]. In the optimum classification, each set of data will have a high percentage of similarity in a certain cluster. Principally, cluster analysis will classify the given set of data that have high similarity in the same cluster and the dissimilar set of data in different Cluster [2].

Clustering is a strategy for arranging patterns (or data items) into groups that do not require supervision (or clusters). The following properties should be present in the partition that is formed as a result: (1) homogeneity within clusters, which means that data from the same cluster should be as comparable as possible. (2) Cluster heterogeneity means that data from various clusters should be as diverse as feasible. Several clustering techniques depend on the number and structure of clusters. Many clustering algorithms that are non-based on GA have acquired extensive usage, including K-means, Fuzzy c means, FSC, etc. [3]. In addition, we don't know how many clusters a particular collection of data consists of the input patterns cannot be automatically and successfully clustered by any of these non-GA-based clustering techniques. This is particularly true when the data set contains a significant number of clusters. In many cases, this was the outcome of a poor initial cluster center selection.

Several clustering algorithms can be optimized by using a genetic algorithm to obtain the optimal solution. Optimization may be applied to parameters or input data.

GAs (Genetic Algorithms) are a kind of evolutionary technique [4]. GAs are used in the process of getting the correct number of combinations and creating acceptable ones. As a result, the parameters of the research were represented as a collection of clustered centroids (a chromosome) [5]. A population consists of chromosomes. In the first step, a representative sample of the research population is chosen at random to reflect the numerous response alternatives. Each chromosome has an objective purpose and a fitness function

associated with it, and these functions combined describe the degree of fitness. Only a small number of chromosomes are selected and passed on to the next generation in line with the theory of "survival of the fittest". Existing chromosomes are exposed to biologically-motivated operators such as crossover and mutation to produce new chromosomes. More and more iterations of selection, crossover, and mutation are carried out in succession until the desired outcome is achieved. The best solution to the clustering problem is found on the chromosome from the preceding generation that was the healthiest and fittest [6]. In this survey, the most important previous studies are presented that used genetic algorithms with several clustering methods, and optimization either on initial value parameters or input data, and show the performance metrics used in each research.

This survey is organized as follows: Section 2 literature review. The genetic algorithms are briefly introduced in Section 3, clustering algorithms are described in Section 4, the role of genetic algorithm in clustering is illustrated in section 5, and some conclusions are shown in Section 6.

II. LITERATURE REVIEW

Most of the previous studies that collected optimization and clustering are old. Therefore, there are recent methods that were used that are not included in the old surveys. So we present a supplement to these surveys that include recent methods.

In [7], this article's major purpose is to present the reader with a thorough and in-depth overview as well as a clear and well-explained notion of a range of approaches and algorithms. The determination of how many clusters are contained in a data user was accomplished in an automated method. Using group-based crossover and a genetic algorithm (GA) to represent the operator, they were able to determine the optimal number of clusters. The researchers broke the enormous project into smaller, more manageable segments. In order to deal with small-scale optimization problems. The assembling process may be improved while simultaneously decreasing the time and complexity required and reducing the overall distance traveled by the salesman. When compared with other approaches that do not integrate GA with k-means clusters, practically all K-means algorithms that employ GA have a high-performance quality of clustering that can be obtained in a short period of time. Using these techniques, the evolutionary process may also arrive at solutions more rapidly.

In [8], this survey used GAs to generate the appropriate number of groups and offer appropriate grouping. Research is being done on a broad range of GA-based clustering algorithms. Some are utilized on very small datasets, while others are put to use on considerably bigger data sets. From production simulation to picture segmentation and clustering to image compression and gene expression analysis to text clustering, genetic algorithms may be used for a wide variety of clustering applications. K-means and fuzzy c-means, which are based on distance, are examples of clustering methods that were adjusted with the aid of GA. As far as they know, no other clustering algorithm has used GA as an input.

In [9], This survey uses a clustering approach called GA-clustering, which was created. The capacity of genetic algorithms to search is employed in order to search for suitable cluster centers in the feature space in such a manner that optimization of a similarity measure for the clusters that are formed may take place. When the chromosomes are shown, they are shown as strings of actual numbers that define where the centers of various clusters are located. It has been extensively shown that the GA-clustering algorithm is superior to the K-means algorithm, which is the algorithm that is most often applied, utilizing five distinct data sets, four of which are false and three of which are real. GAs offer the optimum string when the number of iterations grows to infinity when the chance of traveling from any population to the one with the optimal string is higher than 0. When the number of iterations hits infinity, something occurs. As a result of this, it is envisaged that a GA-based clustering strategy will, given limited conditions, also yield an optimal clustering in regard to the clustering metric that is being researched. Several fake and real-world data sets with the number of dimensions ranging from two to ten and the number of clusters ranging from two to nine have been explored in order to illustrate the usefulness of the GA-based clustering algorithm in providing optimal clustering. From two to nine clusters have been seen. According to the findings, the performance that the GA-clustering algorithm offers for these data sets is noticeably more advantageous than that which is offered by the K-means algorithm.

In [10], presents a GA-based clustering technique. Clustering is widely used in a different fields such as biology, engineering, text mining, bioinformatics, and agriculture. Genetic algorithms more often referred to as GAs, are widely applied to clustering algorithms in order to enhance their efficiency. Most people are acquainted with genetic algorithms as an evolutionary strategy that gives near-global or global optimal solutions for a given fitness function. The capacity of GAs is employed to develop the necessary number of clusters and to give suitable clustering.

III. GENETIC ALGORITHM (GA)

Genetic algorithm, generally known as GAs, are a form of optimization algorithms that carry out adaptive searches in order to disclose answers to large-scale optimization challenges that include numerous local optimal solutions[9]. When cluster analysis is conducted with traditional GAs, chromosomes are utilized to describe the various solution choices. Every chromosome in a cell is a small representation of the full solution. There are two possible results in this situation. In the second case, Cluster analysis of the block was extracted from every location in the solution. GAs are responsible for the control of a large number of chromosomes so that evolution may find the best solution. Through the employment of the tournament and roulette wheel selection operations, as well as the crossover and mutation operators, new offspring are based on the parents of each generation. Crossover and mutation operators are used to create these offspring. Personnel now judged unsuited for their jobs are being phased out and their positions are being replaced by those who have been newly created [4].

GAs are able to escape their local optimal circumstances thanks to the crossover and mutation operators. These operators are highly sensitive to the method in which the likelihood of

crossover and mutation are defined. It has recently been feasible to speed up the genetic processes and break out of local optimal states by adapting operator probabilities in GAs. However, building adaptive GAs is a difficult task, and the vast majority of methods used are heuristic in their approach [11]

IV. CLUSTERING ALGORITHMS

Methods like partitioning and hierarchical structure are utilized in clustering approaches [12]. Aside from (FCM), which employs fuzzy cluster limits, and fuzzy subtractive clustering, which makes use of crisp cluster bounds, there are other partitioning techniques that utilize both fuzzy and crisp cluster bounds. As an example of each, consider the following options: Using fuzzy sets, it is possible to assign each data point to at least one cluster. Therefore, the K-means algorithm cannot function properly unless each data point belongs to precisely one cluster. Consequently, a variety of real-world applications, such as gene expression analysis and pattern recognition, are unlikely to profit from this technology's limitations [13]. The FCM approach, based on fuzzy set theory and allowing each data point to belong to several clusters, is more theoretically sound, but it has become the most often used technique for dividing data sets. However, the FCM algorithm, along with the great majority of other ways of partitioning, is unable to determine the number of clusters completely on its own, and the results are heavily reliant on the beginning parameters. However, the FCM may get stuck in the local best-case scenario for certain initial values [14]. And the subtractive clustering algorithm is mostly based on the density of data points (potential) inside an area. (Variable). The cluster's focus point will be the data point with the greatest potential for utility or importance. The potential value of prospective data points that are placed inside the defined radius of the cluster center will have part of that value deducted. After that, the algorithm will choose a new location to serve as the center of a new cluster. This point will be one that has the potential worth of the highest data point that follows after it. This procedure is carried out again and over again until the requirements that were defined are fulfilled [15].

V. THE ROLE OF THE GENETIC ALGORITHM IN CLUSTERING

The genetic algorithm play important role in the searching capability that is exploited in order to search for appropriate cluster centers in the feature space such that a similarity metric of the resulting clusters is optimized [9]. GA is applied to optimizing input data and parameters in clustering.

A. GA based clustering input data optimization

The In [16], the incremental genetic K-means algorithm (IGKA) was invented by expanding the Fast Genetic K-means Technique, a clustering algorithm previously disclosed (FGKA). When the mutation frequency was low, IGKA outperformed FGKA significantly. The main concept behind IGKA was to calculate the total within-cluster variation (TWCV) objective value and cluster centroids in phases if the mutation chance was low. IGKA has the same distinguishing feature as FGKA in that it always converges to the global optimum. This paper used a time performance measure.

In[17], this paper proposes a fuzzy clustering model to find the proper cluster structures from a dataset used for intrusion detection. Since, a genetic algorithm is an effective technique to improve accuracy. This paper proposed a Novel Weighted Fuzzy C-Means clustering method based on Immune Genetic Algorithm (IGA-NWFCM) and hence it improves the performance of the existing techniques to solve the high dimensional multi-class problems. Moreover, the probability of obtaining the global optimal value is increased by the application of the immune genetic algorithm. This proposed algorithm provides high accuracy, stability, and probability of gaining global optimum value by using precision and recall as performance metrics.

In [18], algorithms used often include clustering by K-Means, fuzzy-C-Means, and subtractive clustering. They advocate the employment of genetic algorithms to sort the data. To test the algorithm's effectiveness, they developed it and ran it on a collection of randomly generated data. They achieved at the conclusion that the performance of the algorithm. When put to broad use, was equivalent to that of the k-means clustering. They made some tweaks to the fitness function and observed that the algorithm functioned as expected and created better clusters than the k-Means clustering algorithm, which was precisely what they wanted to see happen. MSE was used as a performance metric.

In[19] , this paper presents the purpose of optimizing fuzzy c-means clustering, and a kernel-based fuzzy c-means (KFCM) clustering algorithm has been presented. This approach is based on Genetic Algorithm (GA) optimization, which has been combined with an upgraded genetic algorithm and a kernel technique (GAKFCM). The improved adaptive genetic algorithm is used to optimize the first clustering center in this approach. The KFCM method is then used to guide the classification, which helps to improve the FCM algorithm's clustering performance. Both of these algorithms are used to achieve the same goal of optimizing the initial clustering center. The simulation is carried out in the study with the assistance of MATLAB, and the efficacy of the FCM algorithm, the KFCM algorithm, and the GAKFCM algorithm is evaluated with the assistance of test datasets. The findings show that the GAKFCM algorithm presented is capable of effectively overcoming the shortcomings of FCM and significantly enhancing the clustering performance. Precision and recall are used as performance metrics.

In [20], the fundamental objective of this research is to decrease the number of iterative steps necessary for clustering the data to acquire the important information in a shorter amount of time. The technique presently being employed is a genetic algorithm that reduces the overall number of steps. The typical k-means method's step count may be reduced by using GA, as has been found. This paper used mean square error to measure the performance.

In [21], Cluster analysis was done by the use of an evolutionary genetic algorithm (GA) based k-means algorithm that was created in this work. The k-means approach, which has been provided, is used to initiate the GA population. The next step is to apply the GA operators in order to produce a new population. The most distant clustering sites are required for the occurrence of a new mutation. Numerous difficulties may be resolved using the provided technique. Results showed that the newly proposed method is more effective at performing cluster

analyses. The approach that was given is utilized to solve a variety of test questions. The following are some of the many benefits of using the proposed algorithm: There can be no empty clusters in the k-means clustering algorithm since you reject any k-means solutions that contain such empty nodes. Due to the use of a novel mutagenesis technique, the final clusters are devoid of spots determined to be severe. The recommended solution has a speedy convergence, with the number of iterations ranging from 10 to 100 for each problem. The recommended algorithm produces a lower value for the fitness function (SSE) than the k-means algorithm does for the identical issue.

In [22], This study proposes an unsupervised clustering strategy based on a genetic algorithm that seeks out the optimum cluster centers based on a k-means data framework. Using the genetic algorithm, the k-means clustering technique is less delicate and less likely to converge on a local minimum when centers are generated at random. Adding two clustering validity indices to the selection technique makes it possible to automatically determine how many clusters should be produced. Sixteen datasets from diseases and four datasets from single cells are used to demonstrate the usefulness of the algorithm. According to the results, our technique performs substantially better than the algorithms deemed to be highly advanced on the greater part of the datasets. WCSS, DB, and SI were used as evaluation metrics.

B. GA based clustering parameters optimization

In[23], improve the parameters of the subtractive clustering algorithm using an iterative search strategy to offer the optimum number of clusters that the FCM algorithm needs, and then determine the ideal weighting exponent (m) for the FCM algorithm. The settings of the subtractive clustering technique may be adjusted to generate the required number of clusters for the FCM algorithm. The best single-output Sugeno-type Fuzzy Inference System (FIS) model is found by the iterative search. This is accomplished by fine-tuning the subtractive clustering algorithm's parameters to provide the smallest possible difference between the actual data and the Sugeno fuzzy model's predictions. The goal is to have the most clusters possible, and this may be accomplished by repeated searching. To improve the weighting exponent (m) in the FCM algorithm, two approaches are put forward once the number of clusters has been fine-tuned. These approaches are the iterative search methodology and the genetic algorithms. They'll utilize iterative search when they've got the number of clusters to the ideal level of refinement. This approach is tested on data generated by the starting function, and the best fuzzy models are found by finding those with the lowest difference between the real data and the fuzzy models that were derived. This paper used MSE as an evaluation metric.

In[24], this work combines the Fuzzy Clustering Approach with a genetic algorithm (GA), subtractive clustering (SC), and Bayesian cluster validation to build a unique clustering method called fzGASCE. This approach both identifies the proper number of clusters to form a given dataset and does it efficiently. When comparing the results of fzGASCE against those of other strategies that use a similar approach, we find that it performs better. COR, EVAR, and EMIS used as measures of performance.

In [25], The genetic algorithm may employ an objective function to estimate the parameter's optimal value, which demonstrates a simple and solid technique for finding the SC algorithm's parameter ra(cluster radius that is appositive constant defines cluster center range) . To anticipate the optimal value of ra, genetic algorithms (GAs) are utilized and a new objective function is established. CSP measure used in this paper.

In [26], this research, an unusual GA-based clustering technique is provided. This strategy can automatically find the necessary number of clusters and discover the suitable genes by applying a new initial population selection approach. This is accomplished by the creation of high-quality cluster centers employing both our novel fitness function and our gene rearrangement technique. A clustering solution of even higher quality may be developed by allowing the initial seeds, which are employed as seeds in K-Means, to readjust themselves in accordance with the criteria of the clustering process. According to the sign test analysis, the outcomes of our experiments imply that our approach is statistically superior to five more current methods on twenty natural data sets that were applied in this study based on six evaluation criteria such as Xie_Beni (XB), Sum of square error, COSEC, F-measure, Entropy, Purity.

In [27], this study presents FZGPS is a unique algorithm that was built by combining FCM with the Genetic Algorithm (GA), the Particle Swarm Optimization (PSO), and the Subtractive Clustering (SC) techniques. This new algorithm efficiently and successfully decides how many clusters to build from a given dataset. They show that FZGPS operates well on gene expression datasets that are created as well as those that are collected via studies. Accuracy measure used in this paper.

In [28], The method of calculating the parameters k in KNN may affect the accuracy of the classification results. If many parameters are applied, the results of classification are determined by a majority vote. If the KNN classification parameter k was set to 1, the result was highly close since it used the nearest neighbor from the classification results. In contrast, if the KNN parameter k has a big value, the classification results will be unclear. This paper will use the algorithm expectation Maximization (EM) to optimize the parameters k in the UN tax cluster. The results of the study are shown as clustering data using the number of clusters with k value optimization and the number of clusters without k value optimization. After grouping the data, the results are then studied. The findings of the study indicated that k produced by the optimization algorithm may improve cluster outcomes, reducing the 66 % accuracy to 64 % while being extremely close to the best measurement accuracy test result.

In [29], this study presents Automatic genetic fuzzy c-means is a new algorithm that automatically calculates the initial centroids and modifies the number of clusters as the study proceeds. The proposed algorithm is a genetic algorithm, but it incorporates extra operators for crossover, mutation, and modified tournament selection. In addition, three distinct cluster validity indices are employed as the basis for a unique genetic algorithm. To demonstrate the algorithm's utility in terms of data quality, actual data sets are used. Inter cluster distance, intra cluster, and correct ratio are used in this paper.

VI. PERFORMANCE METRICS

Many performance metrics exist, but we will explain the metrics found in each research:

A. MSE (Mean Square Error)

[18][20][23] That can be computed by the equation:

$$\mathrm{MSE} = \sum_{i=1}^{c}\sum_{j=1}^{c1}(|\left|x_i - v_j\right||)^2 \quad (1)$$

Where ||xi- vj || show Euclidean distance linking xi and vj.

C1 number of data points for the i th cluster.

C number of cluster centers.

B. SSE (SUM Of Square Error)

The error in each cluster is defined as the sum of distances between each element and the center of the cluster, therefore the fitness function (Sum of squared error) is defined as the sum of the errors of all clusters [21] that can be computed by the equation:

$$\mathrm{SSE}(C1, C2, \quad ., Ck) = \sum_{i=1}^{k}\sum_{xj\epsilon Ci}\left|\left|x_j - z_i\right|\right| \quad (2)$$

Where *C*1, *C*2… and *Ck* are the clusters to which the data is divided, *Xj* the data belong to the cluster *Ci* and *zi* cluster center of *Ci*.

C. Precision and recall

In [17] [19] computed by applying the two-equation:

$$\mathrm{Precision} = \left[\frac{\mathrm{TP}}{\mathrm{TP+FN}}\right] * 100 \quad (3)$$

$$\mathrm{Recall} = \left[\frac{\mathrm{TP}}{\mathrm{TP+FP)}}\right] * 100 \quad (4)$$

Where TP is a True positive, FN is a false negative, and FP is a False Positive.

D. COSEC, Xie-Beni (XB), and Sum of Square Error (SSE)), and the external (F measure (FM), Entropy (E), and Purity (P)) evaluation criteria [30, 31].

$$\mathrm{XB}(U, Z; X) = \frac{\sigma(U,Z;X)}{n\times \mathrm{sep}(Z)} \quad (5)$$

$$= \frac{\sum_{k=1}^{K}(\sum_{i=1}^{n} u_{ki}^2\, D^2(z_k, x_i))}{n(\min_{k\neq 1}\{D^2(z_k, z_i)\})}$$

Where $\sigma(U, Z; X) = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ki}^2 D^2(z_k, x_i)$

$$\mathrm{sep}(Z) = \min_{k\neq 1}\{D^2(z_k, z_i)\}$$

U, Z, and X represent the partition matrix, set of cluster centers, and the data set, respectively. Note that when the partitioning is compact and good, the value of α should be low while *sep* should be high, there by yielding lower values of *XB* index. The objective is therefore to minimize the *XB* index for achieving proper clustering.

$$\text{F measure (FM)} = 2 \times \frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \quad (6)$$

$$\text{Entropy } (D_i) = -\sum \mathrm{Pr}_i(c_j).\log_2 \mathrm{Pr}_i(c_j)) \quad (7)$$

$$\mathrm{Purity}(D_i) = \max_j(\mathrm{Pr}_j(c_j)) \quad (8)$$

Where $Pr_i(c_j)$ is the proportion of class c_j data points in cluster i orD_i.

E. CSP

For determining the optimum value of ra, this study uses CSP as an objective function for GAs and discovers the lowest value of CSP by GAs with different values of ra. [32]

$$\mathrm{CSP} = \left(\frac{1}{\tau * D_{\min}}\right) * \mathrm{com} \quad (9)$$

Where $D_{\min} = \min_{i\neq j}(||vi - vj||)$

$$\mathrm{com} = \left(\frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{n} \mu^2{}_{ik}\, ||x_i - c_k||^2\right) \quad \text{And } \sum_{k=1}^{c} \mu_{ik} = 1$$

F. COR

Is the correctness ratio [33-36] that computed by equation (10)

$$COR = \frac{1}{n}\sum_{i=1}^{N} I(c - \widehat{c}) \quad (10)$$

Where *N* is several trials, *c* and *c*ˆ are the numbers of clusters and the expected number of clusters, and *I* (.) is defined as:

$$I(X) = \begin{cases} 1, X = 0 \\ 0, X \neq 0 \end{cases}$$

EVAR is a measure of the accuracy of the predicted number of clusters computed by the equation:

$$EVAR = \frac{1}{N}\sqrt{(c - \hat{c})^2} \quad (11)$$

EMIS is the quantity of data items that were improperly classified and forms the foundation for EMIS, which is an assessment of the system's overall performance. Only until an algorithm has calculated the number of clusters in an accurate way can EMIS be computed. After that, the cluster label that was originally allotted to each item is compared with the object's genuine cluster label. There has been a mistake in classification if even one of them does not match [24].

G. Within cluster sum of squares (WCSS)

In [22] within-cluster sum of squares (WCSS): is the objective function of the original k-means. k Refers to the number of clusters, $\{ci; i\epsilon[1..k]\}$ means the cluster centers, and $\{Ci; i\epsilon[1..k]\}$ represent the k clusters (each cluster has many of data sets), the WCSS is defined as:

$$WCSS = \sum_{i=1}^{k}\sum_{j\epsilon c1} ||xj - ci^2|| \quad .. (12)$$

Where xj and center ci are separated by an equation of Euclidean squared distance. The WCSS value of a more effective solution will be lower.

H. Davies and Bouldin (DB) index [37]:

is the function to sum of within-cluster dispersion to between cluster separation be studied here. An improved technique has a lower DB (k) value. To get the DB index, use the formula below:

$$DB(K) = \frac{1}{K}\sum_{i=1}^{k} \max_{i\neq j} (\delta_i + \delta_j/d_{ij}) \qquad (13)$$

Where

K is the number of clusters .

i, j is the ith and jth cluster,

d_{ij} The distance between centers ci and cj,

δ_i and δ_j are the dispersion measure of a cluster Ci and Cj.

For example, Ci represents a cluster's standard deviation of the distance of data points in cluster Ci to the center of this cluster ci.

The silhouette index (SI)[38] compares an item's resemblance to its cluster with the similarity of other clusters (separation). A higher SI value denotes a better level of quality in the solution. The silhouette index is calculated using the following formula:

$$SI = \frac{\sum_{i=1}^{n} \frac{b(i)-a(i)}{max\{a(i);b(i)\}}}{n} \qquad (14)$$

Where

- a(i) $=\frac{\sum_{j\in\{cr\setminus i\}} dij}{n_r-1}$ the average dissimilarity of the ith object, for cluster's objects Cr.
- b(i) =max $\{d_i c_\vartheta\}$, in which diC $=\frac{\sum j\in C_\vartheta d_{ij}}{n_\vartheta}$ the average dissimilarity of ith objects, for cluster objects.

I. SSE

in [29] used a number of metrics such as

_ the number of classes found

– Inter-cluster distance: the distance between the centroids of the clusters. SSE is a commonly used measure. Illustrated in the equation:

$$SSE(C1, C2, \quad ., Ck) = \sum_{i=1}^{k} \sum_{xj\in Ci} ||x_j - z_i||$$

– Intra-cluster distance: the distance between data vectors within a cluster that is computed by three types:

1)Compute diameter distance:computed by

$$d\ (C1,C1)=\max\{d(x_{ips}, x_{jq})\}$$

That is largest distance between an element in one cluster and an element in the other.

2)Average Diameter Distance

$$d\ (C1, C1)=avg\{d(x_{ip}, x_{ja})\}$$

Average distance between elements in one cluster and elements in the other.

3) Centroid Diameter Distance.

$$d\ (C1, C1)=\min\{d(x_{ipr}, x_{jq})\}$$

The smallest distance between an element in one cluster and an element in the other.

VII. CORRECT RATIO RC: THE CORRECT RATIO OF CLUSTER NUMBERS IS DEFINED BY:

$$Rc = TCNC/RT * 100 \qquad (15)$$

Where TCNC represents the number of times when the correct number of clusters was obtained by the algorithm and RT is the number of times where the algorithm was run.

As shown in Table 1, a summary of the most important studies related to different algorithms of clustering for finding the optimal solution by using genetic algorithm optimization

TABLE I. SUMMARY GA BASED CLUSTERING ALGORITHMS

references	Technique	Optimization		Evaluation metrics
		parameters	*Input data*	
[16]	Incremental Genetic K-means Algorithm (IGKA)		√	Time performance
[17]	Ga with FCM		√	Precision, Recall
[18]	Ga +K-Means, FCM, SC		√	MSE
[19]	GA+FCM		√	Precision, recall
[20]	Ga +k_means		√	MSE
[21]	a novel genetic algorithm (GA) based k-means algorithm		√	SSE
[22]	GA-based unsupervised clustering method(Kmean)		√	WCSS,DB,SI
[23]	GA+FCM	√		MSE
[24]	combine FCM with (GA), SC, and Bayesian cluster	√		COR, EVAR, EMIS
[25]	GA+FSC	√		CSP
[26]	Ga +kmean	√		Xie-Beni (XB), SSE, COSEC, FM, E, P
[27]	FCM with (GA), (PSO), and (SC) algorithms	√		Accuracy
[28]	expectation Maximation(EM)	√		Accuracy
[29]	Ga+FCM	√	√	Inter cluster, Intra and Rc

*A. **Conclusion***

Because GAs could be used, they were used to correctly optimize the parameters and cluster input data. Several GA-based clustering algorithms are presently being investigated.

Depending on the size of the data collection, some are applied to smaller datasets, while others are applied to much larger datasets. GA-based clustering algorithms have a broad range of applications. Clustering approaches used in this process include several clustering methods some of them optimized in parameters and another some in input data, which are all mostly based on distance. This study illustrated a range of performance metrics to obtain the optimal results used in each research .this study conclude the most research use accuracy as the fitness function.

REFERENCES

[1] Gan, G., C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. 2020: SIAM.

[2] Rencher, A.C., *A review of "Methods of Multivariate Analysis, "*. 2005, Taylor & Francis.

[3] Gu, L. *A novel locality-sensitive k-means clustering algorithm based on subtractive clustering*. in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. 2016. IEEE.

[4] Katoch, S., S.S. Chauhan, and V. Kumar, *A review on the genetic algorithm: past, present, and future*. Multimedia Tools and Applications, 2021. **80**(5): p. 8091-8126.

[5] Ihsan, M.A., M. Zarlis, and P. Sirait, *Reduction Attributes on K-Nearest Neighbor Algorithm (KNN) using Genetic Algorithm.* 2021.

[6] Tabassum, M. and K. Mathew, *A genetic algorithm analysis towards optimization solutions.* International Journal of Digital Information and Wireless Communications (IJDIWC), 2014. **4**(1): p. 124-142.

[7] Zeebaree, D.Q., et al., *Combination of K-means clustering with Genetic Algorithm: A review*. International Journal of Applied Engineering Research, 2017. **12**(24): p. 14238-14245.

[8] Sheikh, R.H., M.M. Raghuwanshi, and A.N. Jaiswal, *Genetic Algorithm Based Clustering: A Survey*. 2008: p. 314-319.

[9] Maulik, U. and S. Bandyopadhyay, *Genetic algorithm-based clustering technique.* Pattern recognition, 2000. **33**(9): p. 1455-1465.

[10] Sumbherwal, N. and B. Hooda, *Genetic algorithm-based clustering methods: A review*. 2021.

[11] Almadhoun, W. and M. Hamdan *Optimizing the self-organizing team size using a genetic algorithm in agile practices.* Journal of Intelligent Systems, 2020. **29**(1): p. 1151-1165.

[12] Haryati, A.E. and S. Surono, *COMPARATIVE STUDY OF DISTANCE MEASURES ON FUZZY SUBTRACTIVE CLUSTERING*. MEDIA STATISTIKA.2020. **14**(2): p. 137-145.

[13] Kour, H., J. Manhas, and V. Sharma, *Evaluation of Subtractive Clustering-based Adaptive Neuro-Fuzzy Inference System with Fuzzy C-Means based ANFIS System in Diagnosis of Alzheimer*. Journal of Multimedia Information System, 2019. **6**(2): p. 87-90.

[14] Lambora, A., K. Gupta, and K. Chopra. *Genetic algorithm-A literature review*. in *2019 international conference on machine learning, big data, cloud, and parallel computing (COMITCon)*. 2019. IEEE.

[15] Irwandi, O.S. Sitompul, and R.W. Sembiring, *Performance Analysis of Subtractive Clustering Algorithm in Determining the Number and Position of Cluster Centers.* Randwick International of Social Science Journal, 2021. **2**(2): p. 193-199.

[16] Lu, Y., et al., *Incremental genetic K-means algorithm and its application in gene expression data analysis.* BMC bioinformatics, 2004. **5**(1): p. 1-10.

[17] Ganapathy, S., et al., *A novel weighted fuzzy C–means clustering based on immune genetic algorithm for intrusion detection.* Procedia Engineering, 2012. **38**: p. 1750-1757.

[18] Kala, R., A. Shukla, and R. Tiwary, *A novel approach to clustering using genetic algorithm.* International Journal of Engineering Research and Industrial Applications, 2010. **3**(1): p. 81-88.

[19] Ding, Y. and X. Fu, *Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm.* Neurocomputing, 2016. **188**: p. 233-238.

[20] Irfan, S., G. Dwivedi, and S. Ghosh. *Optimization of K-means clustering using genetic algorithm*. in *2017 International conference on computing and communication technologies for the smart nation (IC3TSN)*. 2017. IEEE.

[21] El-Shorbagy, M.A., et al. *A novel genetic algorithm based k-means algorithm for cluster analysis*. in *International Conference on Advanced Machine Learning Technologies and Applications*. 2018. Springer.

[22] Nguyen, H., S.J. Louis, and T. Nguyen. *MGKA: A genetic algorithm-based clustering technique for genomic data*. in *2019 IEEE Congress on Evolutionary Computation (CEC)*. 2019. IEEE.

[23] Alata, M., M. Molhim, and A. Ramini, *Optimizing of fuzzy c-means clustering algorithm using GA.* Update, 2008. **1**(5).

[24] Le, T., T. Altman, and K.J. Gardiner. *A fuzzy clustering method using Genetic Algorithm and Fuzzy Subtractive Clustering*. in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*. 2012. The Steering Committee of The World Congress in Computer Science, Computer ….

[25] Shieh, H.-L., P.-L. Chang, and C.-N. Lee. *An efficient method for estimating cluster radius of subtractive clustering based on a genetic algorithm*. in *2013 IEEE International Symposium on Consumer Electronics (ISCE)*. 2013. IEEE.

[26] Rahman, M., A., and M.Z. Islam, *A hybrid clustering technique combining a novel genetic algorithm with K-Means.* Knowledge-Based Systems, 2014. **71**: p. 345-365.

[27] Le, T. and L. Vu. *A novel fuzzy clustering method based on GA, PSO, and Subtractive Clustering*. in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. 2020. IEEE.

[28] Lubis, Z., P. Sihombing, and H. Mawengkang. *Optimization of K Value at the K-NN algorithm in clustering using the expectation-maximization algorithm*. in *IOP Conference Series: Materials Science and Engineering*. 2020. IOP Publishing.

[29] Jebari, K., A. Elmoujahid, and A. Ettouhami, *Automatic genetic fuzzy c-means.* Journal of Intelligent Systems, 2020. **29**(1): p. 529-539.

[30] Larose, D.T., *An introduction to data mining.* Traduction et adaptation de Thierry Vallaud, 2005.

[31] Mukhopadhyay, A. and U. Maulik, *Towards improving fuzzy clustering using support vector machine: Application to gene expression data.* Pattern Recognition, 2009. **42**(11): p. 2744-2763.

[32] Zanaganeh, M., S.J. Mousavi, and A.F.E. Shahidi, *A hybrid genetic algorithm–adaptive network-based fuzzy inference system in prediction of wave parameters.* Engineering Applications of Artificial Intelligence, 2009. **22**(8): p. 1194-1202.

[33] Halder, A., S. Pramanik, and A. Kar, *Dynamic image segmentation using fuzzy c-means based genetic algorithm.* International Journal of Computer Applications, 2011. **28**(6): p. 15-20.

[34] Ghosh, A., N.S. Mishra, and S. Ghosh, *Fuzzy clustering algorithms for unsupervised change detection in remote sensing images.* Information Sciences, 2011. **181**(4): p. 699-715.

[35] Lianjiang, Z., Q. Shouning, and D. Tao. *Adaptive fuzzy clustering based on genetic algorithm*. in *2010 2nd International Conference on Advanced Computer Control*. 2010. IEEE.

[36] Liu, Y. and Y. Zhang. *Optimizing parameters of the fuzzy c-means clustering algorithm*. in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*. 2007. IEEE.

[37] Davies, D. and D. Bouldin, *A cluster separation measure, IEEE transactions on patter analysis and machine intelligence. vol*. 1979, PAMI-1.

[38] Rousseeuw, P.J., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics, 1987. **20**: p. 53-65.